trcls: Classifying SAM Alignments by Splice Variant

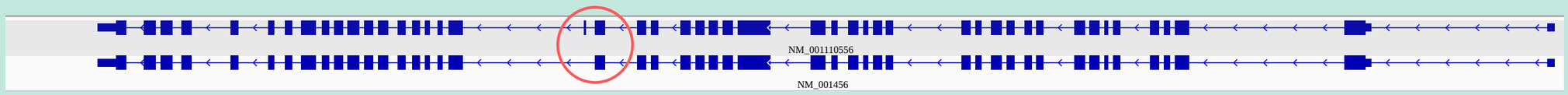
Ning Yuan LEE, ningyuan@u.nus.edu

Abstract

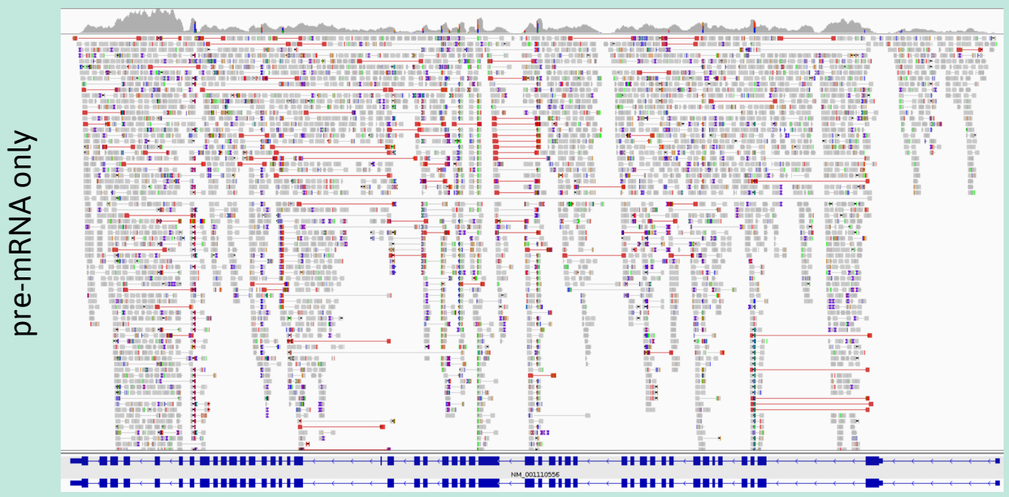
Two sequence reads aligning to the same chromosomal region may represent different RNA species and therefore possess different biochemical characteristics, if they are spliced differently. While there are many pre-existing tools for the quantification of splice-variants from raw or pre-aligned reads, none seem to support the mere classification of pre-aligned reads to facilitate further isoform-specific downstream analysis. To fill this need, a command line tool *trcls* was developed which annotates a SAM alignment file with each read's potential isoform of origin, given a GTF annotation file.

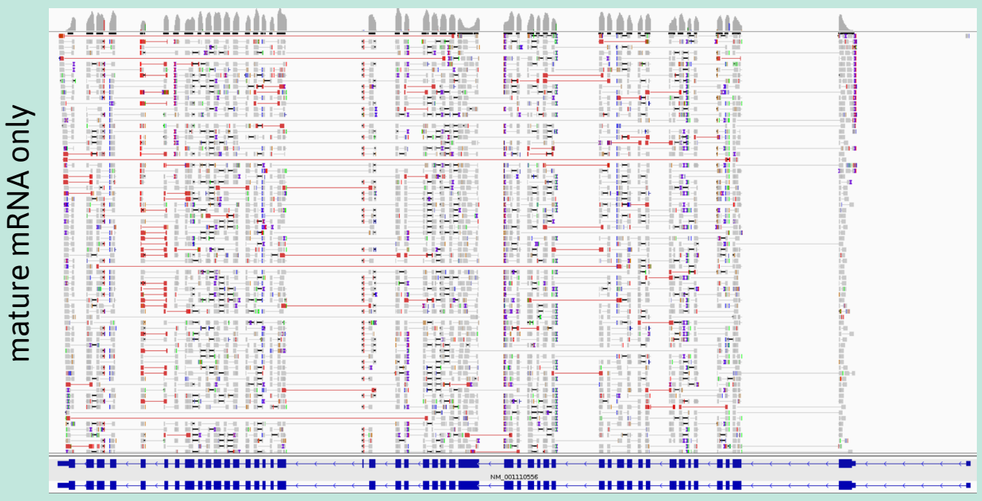
Results

The efficacy of trcls was assessed by its ability to classify mRNA splice variants of human filament A, alpha (FLNA). Two mRNA splice variants are known, dubbed here variant A and variant B. ***Variant A has one extra exon which variant B skips.***

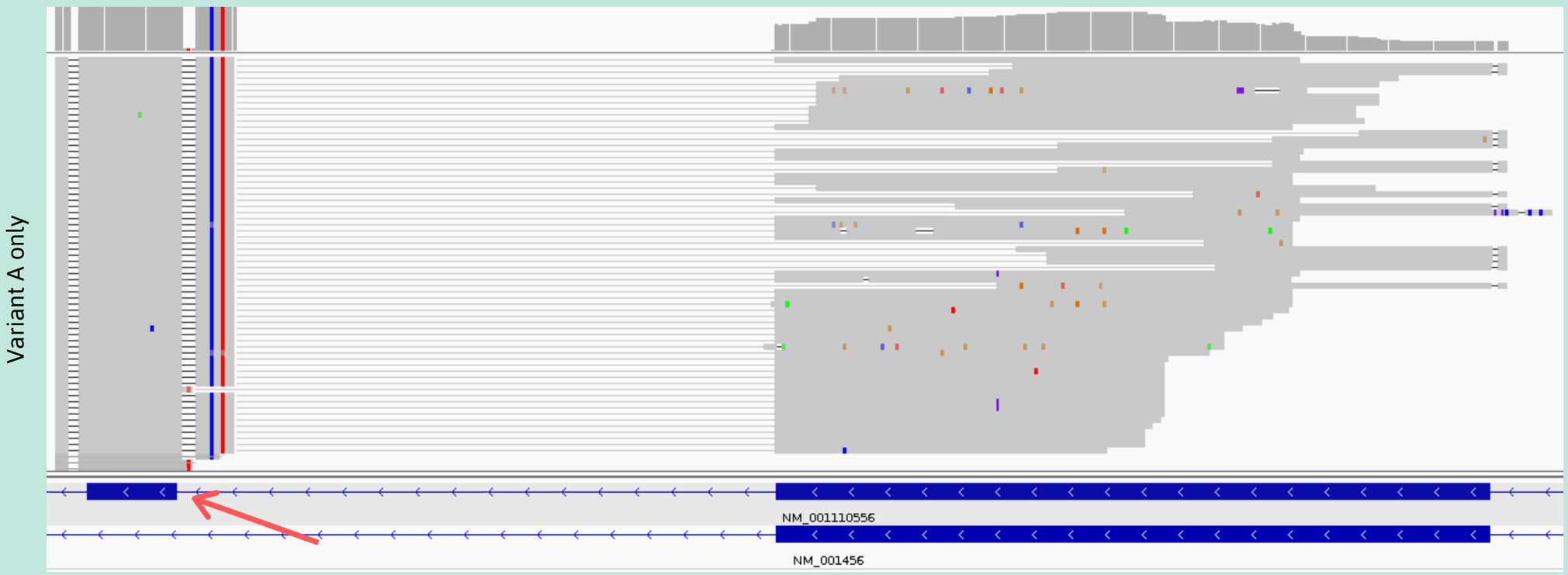


Firstly, it can be observed via simple inspection that trcls is able to classify pre-mRNA from mature mRNA:

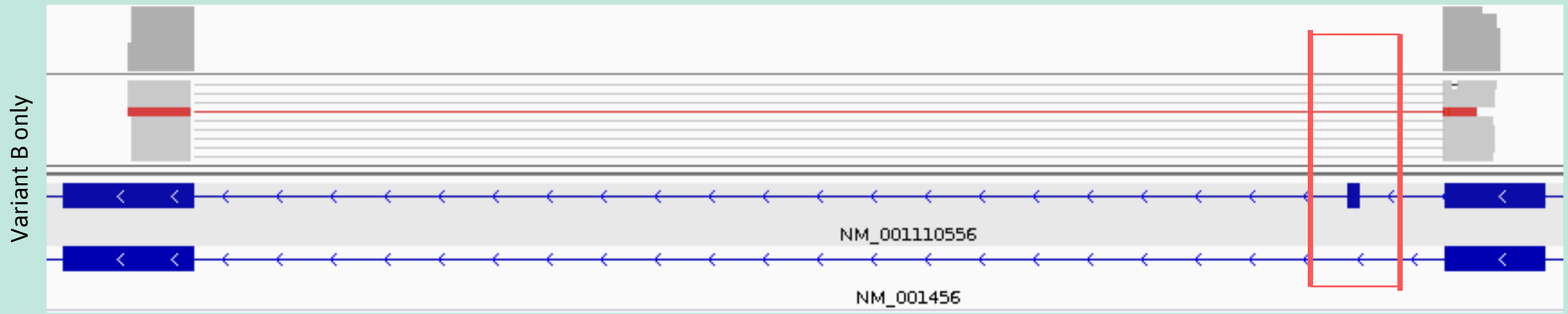




Next, it can be shown empircally that trcls identifies variant A as expected, form its extra exon:



And variant B, by its exon skip:



Conclusion

The isolation of specific splice variants from aligned sequences may further ease the study of RNA biology at isoform-resolution. By examining alignments failing the annotation process---those that do not seem to originate from known isoforms---it is also possible to discover new splice variants.